Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Non-Locality in Chemical Space

Journal:	*The Journal of Physical Chemistry Letters*
Manuscript ID:	jz-2015-008319.R2
Manuscript Type:	Letter
Date Submitted by the Author:	04-Jun-2015
Complete List of Authors:	Hansen, Katja; Fritz-Haber-Institute, Theory Biegler, Franziska; Technical University of Berlin, Machine Learning Group Ramakrishnan, Raghunathan; University of Basel, Chemistry Pronobis, Wiktor; Technical University of Berlin, Machine Learning Group von Lilienfeld, O. Anatole; University of Basel, Institute of Physical Chemistry Müller, Klaus-Robert; Technical University of Berlin, Machine Learning Tkatchenko, Alexandre; Fritz-Haber-Institut der Max-Planck-Gesellschaft,

Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Non-Locality in Chemical Space

Katja Hansen[1], Franziska Biegler[2], Raghunathan Ramakrishnan[3], Wiktor Pronobis[2], O. Anatole von Lilienfeld[3,4], Klaus-Robert Müller[2,5,*] and Alexandre Tkatchenko[1*]

[1] *Fritz-Haber-Institut der Max-Planck-Gesellschaft, Faradayweg 4-6, 14195, Berlin, Germany*

[2] *Machine Learning Group, Technical University of Berlin, Marchstr. 23, 10587 Berlin, Germany*

[3] *Institute of Physical Chemistry and National Center for Computational Design and Discovery of Novel Materials, Department of Chemistry, University of Basel, Klingelbergstrasse 80, CH-4056 Basel, Switzerland*

[4] *Argonne Leadership Computing Facility, Argonne National Laboratory, Argonne, Illinois 60439, USA*

[5] *Department of Brain and Cognitive Engineering, Korea University, Korea*

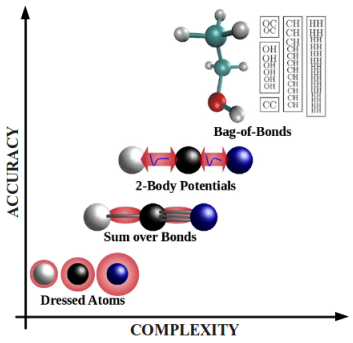
E-mail: klaus-robert.mueller@tu-berlin.de; tkatchenko@fhi-berlin.mpg.de

Abstract

Simultaneously accurate and efficient prediction of molecular properties throughout chemical compound space is a critical ingredient toward rational compound design in chemical and pharmaceutical industries. Aiming towards this goal, we develop and apply a systematic hierarchy of efficient empirical methods to estimate atomization and total energies of molecules. These methods range from a simple sum over atoms, addition of bond energies, pairwise interatomic force fields, reaching to the more sophisticated machine learning approaches that are capable of describing collective interactions between many atoms or bonds. In the case of equilibrium molecular geometries, even simple pairwise force fields demonstrate prediction accuracy comparable to benchmark energies calculated using density-functional theory with hybrid exchange-correlation functionals. However, accounting for the collective many-body interactions proves to be essential for approaching the "holy grail" of chemical accuracy of 1 kcal/mol for both equilibrium and out-of-equilibrium geometries. This remarkable accuracy is achieved by a vectorized representation of molecules (so-called *Bag-of-Bonds* model) that exhibits strong non-locality in chemical space. In addition, the same representation allows us to predict accurate electronic properties of molecules, such as their polarizability and molecular frontier orbital energies.

Keywords

chemical compound space — machine learning — atomization energies — molecular properties — many-body potentials



Chemical compound space (CCS) is the space populated by all possible energetically stable molecules varying in composition, size, and structure.[1] Chemical reactions and transformations due to external perturbations allow us to explore this astronomically large space in order to obtain molecules with desired properties (e.g., stability, mechanical, and electronic properties). The accurate prediction of these molecular properties in the CCS is a critical ingredient toward rational compound design in chemical and pharmaceutical industries. Therefore, one of the major challenges is to enable quantitative calculations of molecular properties in CCS at moderate computational cost (milliseconds per molecule or faster). However, currently only wavefunction-based quantum-chemical calculations, which can take up to several days per molecule, consistently yield the desired "chemical accuracy" of 1 kcal/mol required for predictive *in silico* rational molecular design.

Leaving aside the quest for accuracy, even our understanding of the structure and properties of CCS is remarkably shallow. Furthermore, a unique mathematical definition of CCS is lacking because the mapping between molecular geometries and molecular properties is often not unique, meaning that there can be structurally different molecules exhibiting very similar values for any given property. This complexity is reflected by the existence of hundreds of descriptors that aim to measure molecular similarity in chemoinformatics.[2,3] In this context, one of our goals is to shed light into the structure and properties of CCS in terms of molecular atomization energies that is an essential molecular property measuring the stability of a molecule with respect to its constituent atoms. Atomization energies are accessible exper-

imentally and are frequently used to assess the performance of electronic structure methods. The total energy of a molecule can be trivially determined from its atomization energy by simply adding free atom energies. Under certain conditions chemical reaction barriers can also be correlated to the difference between total energies of two molecules. Obviously, total energies are insufficient to predict the stability and reactivity of molecules in realistic environments, therefore we will have to eventually account for thermodynamic and kinetic effects beyond total energies.

In this Letter, we gradually construct more reliable models that include one-, two- and finally many-body interactions between atoms or bonds. This cascade of models highlights the importance of many-body effects and also illustrates to which amount they can be incorporated as effective terms of lower complexity. Moreover, the *Bag-of-Bonds* approach introduced here enables us to demonstrate the impact of non-local information in CCS that turns out to be crucial for achieving a prediction accuracy of 1.5 kcal/mol for a database of more than 7000 organic molecules. Our research is aimed towards the goal of understanding the structure and properties of CCS composed of molecules with arbitrary stoichiometry. We would therefore like to distinguish our work on predicting molecular properties for varying molecular composition from other very important efforts on constructing potential-energy surfaces ("force fields") of molecules and solids.[4–8] Molecules at their equilibrium structures form a well-defined submanifold of the CCS, and in this manuscript we focus on the properties of molecules in this fundamental submanifold.

Evidently, the dimensionality of CCS grows exponentially with increasing molecular size. However, typical databases of synthetically accessible molecules are rather restricted in their composition. To avoid systematic bias yet enable complete exploration of a subset of CCS, we selected all 7165 molecules from the GDB database containing up to seven "heavy" (C, N, O, S) atoms saturated with hydrogens to satisfy valence rules[9,10] (this database is referred to as GDB-7 throughout this work). In contrast to other widely employed databases, GDB includes *all* molecular graphs corresponding to a set of simple organic chemistry rules.

Atomization energies range from -800 to -2000 kcal/mol. Structural features include a rich variety of chemistry such as double and triple bonds, (hetero)cycles, carboxy, cyanide, amide, alcohol, and epoxy groups. For each of the many stoichiometries, many constitutional (differing chemical bonds) but no conformational isomers are part of this database. The initial geometries in the GDB-7 database were generated using OpenBabel from their associated SMILES descriptors.[11] Subsequently, the geometries were optimized using density functional theory calculations with the PBE exchange-correlation functional[12] in the FHI-aims code.[13] Finally, the atomization energies were computed using the hybrid PBE0 functional[14] at the optimized PBE geometries. The PBE0 functional yields atomization energies with an overall accuracy better than 5 kcal/mol when compared with experiment,[15] and these atomization energies will be used as a reference.

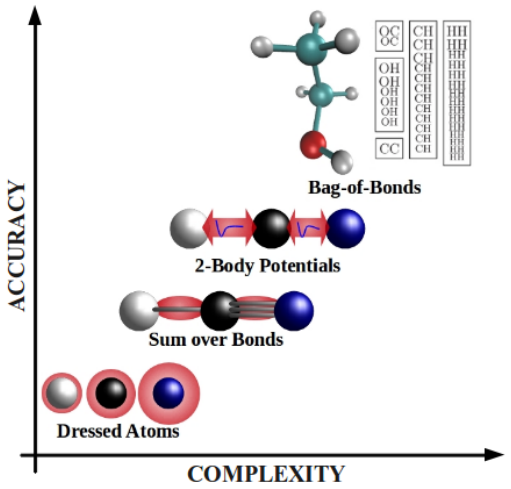


Figure 1: Schematic overview of employed modeling approaches: The dressed atoms model incorporates only the atoms and weights them according to their type. The Sum-over-Bonds and 2-Body potentials consider pairs of atoms and the interactions between them. The Bag-of-Bonds model, which is illustrated for the ethanol molecule (C_2H_5OH) implements a collective energy expression based on all interatomic distances within a molecule.

The hierarchy of models that we use to predict atomization energies is schematically illustrated in Figure 1, ordered from crudest models to the more sophisticated ones in order of complexity. The performance of the different models, evaluated using a rigorous five-fold cross-validation approach,[16] is shown in Table 1.

In the most naive approximation, the energy of a molecule $\mathbf{M}$ can be written as a sum of effective atomic contributions $\hat{E}_{\mathrm{DA}}(\mathbf{M}) = \sum_\alpha n_\alpha E_\alpha$, where $\hat{E}_{\mathrm{DA}}$ refers to the approximated atomization energy, E_α is a "dressed" atomic energy for atom type α (with $\alpha \in \mathrm{H, C, N, O, S}$), and n_α is the number of α-type atoms. This atomic model yields an accuracy of 15.1 kcal/mol on the GDB-7 database (see Table 1).

Table 1: Performance of different models evaluated out-of-sample in five-fold cross-validation on the GDB-7 database: The data was randomly split into five sets of 1433 molecules each. Each of these five sets serves once as test set while the remaining 5732 molecules are used for training. The performance of the models averaged over the five runs is shown, as measured by the mean absolute error (MAE) on the test set (with variance below 3% for all models).

Model	MAE [kcal/mol]
Dressed atoms	15.1
Sum-over-bonds	9.9
Lennard-Jones potential	8.7
Polynomial pot. ($n = 6$)	5.6
Polynomial pot. ($n = 10$)	3.9
Polynomial pot. ($n = 18$)	3.0
Bag-of-Bonds ($p = 2$, Gaussian)	4.5
Bag-of-Bonds ($p = 1$, Laplacian)	1.5
Coulomb matrix ($p = 2$, Gaussian)[17]	10.0
Coulomb matrix ($p = 1$, Laplacian)[16]	4.3

Molecules form as a result of chemical bonding, hence an approach that considers bonds rather than atoms is expected to perform much better. We define a bond by the type of covalently bonded atoms (C,N,O,S) and bond order (single, double, triple), and compute the energy as a sum over all bonds in the molecule. The bond energies are fitted on the GDB-7 database. This definition leads to the so-called *sum-over-bonds* model, which improves significantly over the atomic model, achieving an accuracy of 9.9 kcal/mol on the GDB-7 database. However, the sum-over-bonds model is still unable to treat changes in bond distances and interatomic interactions beyond nearest neighbors.

Both of these effects can be included by constructing an effective pairwise interatomic potential

$$\hat{E}_{\mathrm{PP}}(\mathbf{M}) = \sum_{ij}^{\text{Atom pairs}} \sum_{r}^{r \text{ of type } ij} \Phi_{ij}(r), \quad (1)$$

where $\Phi(r)$ is an effective potential function for each type of atom pair ij (carbon–carbon C–C, carbon–nitrogen C–N and so on). We note that different functional forms can be adopted for $\Phi(r)$, ranging from Lennard-Jones and Morse-type to more general polynomial potentials. Already the usage of Lennard-Jones potential yields an accuracy of 8.7 kcal/mol on the GDB-7 database. This is because the Lennard-Jones potential reproduces the basic features which a general interatomic potential should possess – repulsive wall at short distances, a well-defined minimum and the van der Waals r^{-6} decay of the interaction at large interatomic distances.

More general pairwise potentials can be constructed by a systematic expansion of $\Phi(r)$ using powers of the inverse distance r^{-n}. The performance of such polynomial models as a function of the maximum degree n is shown in Table 1. The improvement in atomization energies saturates around $n = 18$, reaching an accuracy of 3.0 kcal/mol. To put this number in perspective, we recall that 3 kcal/mol is below the error of the reference PBE0 atomization energies when compared to experiment.[15] Moreover, the performance of the most sophisticated machine learning (ML) models in Ref.[16] applied to a similar dataset was 3.1 kcal/mol. Therefore, it is remarkable that a simple and very efficient model based on pairwise potentials is able to capture the subtle energetic contributions required to predict atomization energies for equilibrium molecular geometries.

Seeking to better understand this finding, we plot the optimized C–C potential in Figure 2 for different values of n. The increase in the degree of the polynomial leads to the appearance of shoulders and minima related to different bond orders. In fact, these features appear at interatomic distances well known from empirical determinations of bond orders and energies.[18] We thus conclude that the increase in the degree of the polynomial enables

the potential to "learn" about chemical bonding. Similar observations as for the C–C potential are demonstrated for C–N and and C–O potentials in the supplemental material. The improvement in the predictive power of polynomial potentials does not only arise from their ability to distinguish between different bonding scenarios. The decay of these potentials with interatomic distance is rather slow, with energy contributions beyond nearest neighbors (> 1.5 Å) having an essential role on the scale of the obtained error (see inset in Fig. 2). We note in passing that another attractive feature of interatomic potentials is that by construction they can exactly reproduce the limit of dissociated atoms, a condition that is difficult to fulfill even in state-of-the-art *ab initio* theory. While we used polynomial potentials in this work, other choices of basis functions are certainly possible, but no significant accuracy gains are found, e.g., utilizing spline-based potentials.

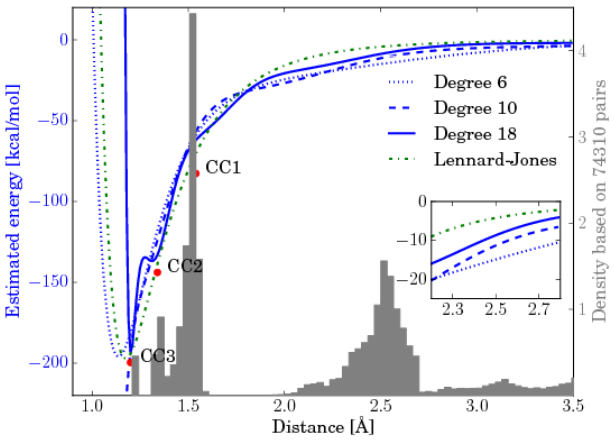


Figure 2: Polynomial potentials for C–C interaction: The normalized gray histogramm refers to the distribution of C–C distances within the GDB-7 dataset and is associated with the right-hand axis. The red dots represent the energies of the C–C single, double and triple bond as given by fits to experimental bond energies.[18] In blue, polynomial two-body potentials (as trained in cross validation) are shown. The inset shows the difference between potentials for distances between 2.2 Å and 2.8 Å.

While the performance of pairwise potentials is already quite good, they have a few notable drawbacks. For example, their performance for out-of-equilibrium molecular geometries is strongly degraded. In order to demonstrate this, we extended the GDB-7 database by scaling all the interatomic distances in the molecules by a factor of 0.9 and 1.1. When

trying to learn the atomization energies for out-of-equilibrium molecular geometries, the performance of pairwise potentials diminished by 16.7 kcal/mol compared to pure equilibrium geometries. This test demonstrates that while pairwise potentials can be successfully applied in preliminary studies of stabilities for equilibrium geometries (when these are given from some other method), more sophisticated approaches are required for non-equilibrium molecular geometries.

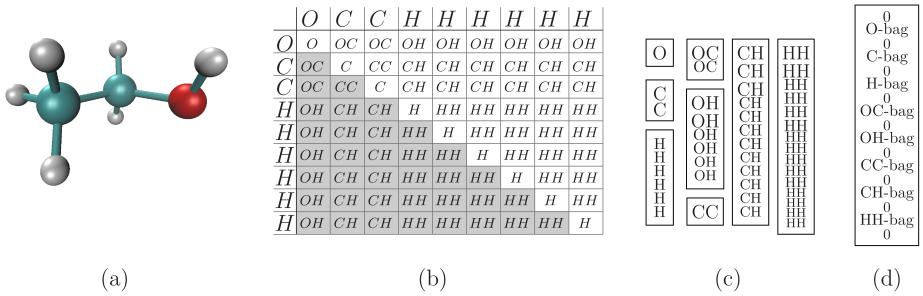


Figure 3: Schematic view of the Bag-of-Bonds (BoB) representation. (a) Shows the three-dimensional structure of ethanol (CH_3CH_2OH) and (b) specifies the involved nuclear charges for each Coulomb matrix element. In (c) the different Coulomb matrix entries which are present for ethanol are sorted into bags and the BoB vector (d) is obtained by concatenating these bags and adding zeros to allow for dealing with other molecules with larger bags.

Evidently, collective effects beyond pairwise potentials are important for chemically accurate modeling of molecular atomization energies. To include these effects, we propose a more sophisticated ML approach, which we call *Bag-of-Bonds* (BoB). The BoB concept is inspired by text mining descriptors utilized in computer science[19,20] (see Figure 3 and supplemental material for a detailed description of the model). In natural language processing, the so-called bag-of-words descriptor that encodes the frequency of occurence of words in text is used for solving classification problems.[19,20] Here, instead we propose to use interatomic (inverse) distances in the BoB descriptor for accurate predictions throughout chemical compound space. In the BoB model, first the molecular Hamiltonian is mapped to a well-defined descriptor, here a vector composed of *bags*, where each bag represents a particular bond type (C–C, C–N and so on). Motivated by the Coulomb matrix concept of Rupp *et al.*,[17] each entry in every bag is computed as $Z_i Z_j / |\mathbf{R}_i - \mathbf{R}_j|$, where Z_i and Z_j are

the nuclear charges while $\mathbf{R}_i$ and $\mathbf{R}_j$ are the positions of the two atoms participating in a given bond. In order to vectorize this information, instead of forming a matrix we simply concatenate all bags of bonds in a specified order (the order is irrelevant for the learning process), padding each bag with zeros in order to give the bags equal sizes across all molecules in the GDB-7 database and sorting the entries in each bag according to their magnitude. This representation is naturally invariant under molecular rotations and translations, whereas the permutational invariance is enforced by the sorting step. We note in passing that unlike the sorted Coulomb matrix[17] the BoB descriptor is not able to distinguish between homometric molecules[21] (molecules with different geometries, but equal set of pairwise distances between nuclei), however our database is devoid of such cases.

We split the full GDB-7 database into a training set of N molecules and a testing set containing the rest of the molecules (cf.[16]). The energy of a molecule with a BoB vector $\mathbf{M}$ is written as a sum over weighted exponentials centered on every molecule I in the training set

$$\hat{E}_{\mathrm{BoB}}(\mathbf{M}) = \sum_{I=1}^{N} \alpha_I \exp\left(-d(\mathbf{M}, \mathbf{M}_I)/\sigma\right), \tag{2}$$

where $d(\mathbf{M}, \mathbf{M}_I) = \sum_j \|M^j - M_I^j\|_p$ defines the distance (not necessarily Cartesian) between the BoB vectors $\mathbf{M}$ and $\mathbf{M}_I$ ($\|x\|_p$ refers to the l_p norm of x), α_I are the regression coefficients, the kernel width σ is optimized for each choice of p by five-fold cross-validation,[16] and I runs over all molecules $\mathbf{M}_I$ in the training set of size N. The values of α_I coefficients and σ are determined by a kernel-ridge regression (KRR) procedure as described in detail elsewhere.[16,17] KRR is a standard robust technique in machine learning which limits the norm of regression coefficients, α_I, thereby ensuring the transferability of the BoB model to new compounds.

To understand the physics behind the BoB model, we can decompose the BoB Laplacian kernel for a molecule $\mathbf{M}$ as $\exp(-\sum_j^n |M^j - M_I^j|/\sigma) = \prod_j^n \exp(-|M^j - M_I^j|/\sigma)$. Taylor-series expansion of the exponential as a function of internuclear Coulomb repulsion and the subsequent product will include contributions up to infinite order in terms of bond pairs

between molecules $\mathbf{M}$ and $\mathbf{M}_I$. We stress that the BoB model uses implicitly the same ingredients as conventional multipolar potentials, albeit with a different, arguably more general, functional form. Simple sum over bonds and pairwise potential approaches can be constructed as lower-order expansions of the BoB model, given sufficient training data. In fact, a connection between the BoB model and pairwise potentials can be established by approximately rewriting the BoB kernel as $\sum_l^{n_b} \prod_{j \in b_l} \exp(-|M^j - M_I^j|/\sigma)$, where b refers to a certain type of bond (e.g., C–C) and n_b is the length of the bag corresponding to the bond type b. We found that such partial linearization of the BoB model reduces the accuracy, reverting the performance back to the pairwise polynomial potential model. This clearly demonstrates the crucial role of collective many bond effects accounted for by the non-linear infinite-order nature of the kernel. We note that Eq. 2 only includes contributions from *pairs* of molecules. One could also envision more complex approaches that correlate information from three or more molecules at a time.

The flexibility in choosing the kernel metric in CCS (the function d in Eq. 2) allows us to investigate the locality properties of chemical space for the prediction model in terms of atomization energies. The high sensitivity of the BoB model on the employed kernel is demonstrated in Table 1, where a more local (in terms of distance in chemical space) Gaussian kernel ($p = 2$) leads to an accuracy of 4.5 kcal/mol versus a much improved performance of 1.5 kcal/mol for a non-local Laplacian kernel ($p = 1$). We remark that the remarkable performance of the BoB model with the Laplacian kernel with respect to previous work[16,17] is far from being a trivial achievement. In the context of standard quantum-chemical calculations, the improvement of accuracy from 3.1 kcal/mol[16] to 1.5 kcal/mol would imply an increase of several orders of magnitude in the computational cost. However, the cost of BoB calculations is the same as that of the previous less accurate ML methods in Refs.[16,17] Hence, the development of the BoB model takes machine learning approaches to an unprecedented level of accuracy, enabling calculations close to the "holy grail" of chemical accuracy for equilibrium molecular geometries throughout chemical compound space. We

note in passing that determining equilibrium molecular geometries as an input for BoB calculations is not a difficult task, and even simple and efficient semi-empirical quantum-chemical approaches yield accurate results for equilibrium molecular geometries.

To further elucidate the role of non-local information in chemical space in the prediction of atomization energies, we have systematically studied the dependence of the prediction accuracy on the metric norm p employed in Eq. 2. We find that the optimal value of p is close to unity and the predictive capability decreases significantly for $p < 0.5$ and $p > 1.5$. For larger values of p, e.g. $p = 2$, the resulting model is more local and yields worse results. For kernel-based models it is possible to calculate the contribution to the predicted value for each compound in the training set. Adding up all contributions from compounds close to the compound in question we obtain a "local estimate" of the predicted value. Figure 4 illustrates how this local estimate of the atomization energy converges towards the predicted value with randomly selected and growing molecular neighborhoods in the case of ethanol molecule for Gaussian and Laplacian kernels. Clearly, the Laplacian kernel is able to optimally utilize non-local information in CCS. This is further demonstrated by analyzing the optimized kernel width σ in Eq. 2 corresponding to the Gaussian and Laplacian kernels. The value of σ fluctuates widely for the Gaussian kernel for different training set sizes in Figure 4 as does the standard deviation when training on independently drawn training sets. The corresponding fluctuations are smaller for the Laplacian kernel and σ reaches its converged value after $N = 500$. Similar results as for ethanol are found for the other molecules in the GDB-7 database. The issue of (non)locality in the accurate prediction of molecular properties leads to the question of whether it is possible to identify a minimal set of molecular structural fragments which would be sufficient to preserve the good accuracy of the BoB model. Such finding would allow us to extend the applicability of the BoB model to much larger molecules, and this will be a subject of our future work.

In contrast to pairwise potentials, the good performance of the BoB approach extends also to non-equilibrium molecular geometries. For the extended GDB-7 database with stretched

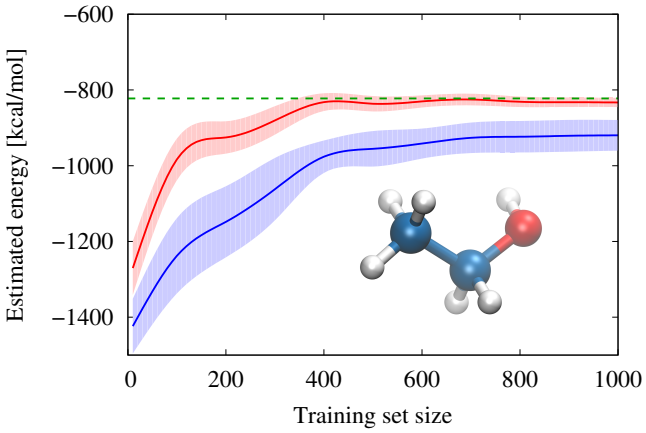


Figure 4: Estimated atomization energy of the ethanol molecule (C_2H_5OH) as predicted by the BoB model using Gaussian (blue line) and Laplacian (red line) kernels. The PBE0 reference energy is indicated by the dashed green line. For a given training set size, the estimation is an average of predictions from 10 optimized models, each employing independently sampled training molecules (excluding ethanol) from the GDB-7 database. The envelope encloses the standard deviation of the estimate from 10 independent runs.

and compressed geometries described above, the prediction error of BoB increases only by 0.8 kcal/mol. This is a direct reflection of the ability of the BoB approach to correctly capture the intricate collective interactions between many bonds within organic molecules.

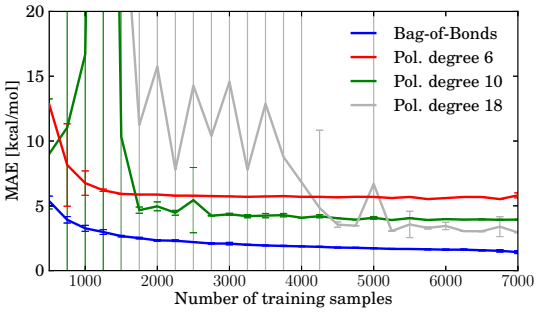


Figure 5: Mean absolute error (MAE in kcal/mol) for BoB and polynomial models: Training sets from $N = 500$ to 7000 data points were sampled identically for the different methods. The polynomial model of degree 10 and 18 exhibit high variances due to the random stratification, which for small N leads to non-robust fits.

Another advantage of the BoB model over pairwise potentials is its better transferability and smooth prediction improvement with the number of training samples, as shown in Figure 5. Already when using just 1000 random molecules out of GDB-7 for training, the

BoB model demonstrates prediction accuracy comparable to the best optimized polynomial potential with degree 18, which requires more than 5000 training samples to achieve the same level of accuracy. At a first glance, this is surprising considering that the polynomial potential contains less adjustable parameters. However, Figure 5 demonstrates that BoB represents a more robust machine learning model with proper regularization and that further improvement in accuracy is possible by simply enlarging the molecular database. This demonstrates the great promise of the BoB approach for further exploration and understanding of CCS.

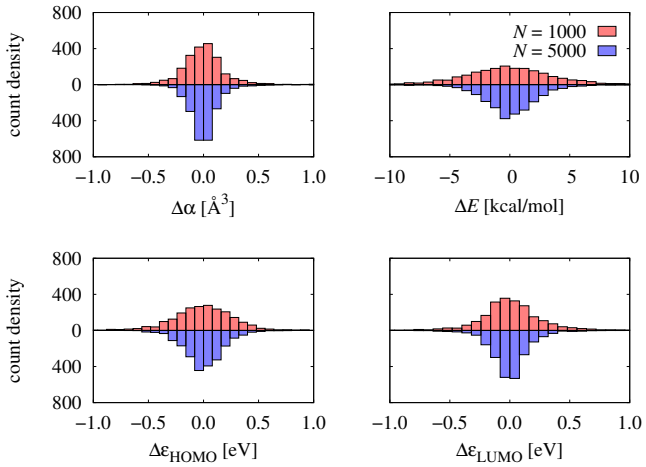


Figure 6: Error distribution of BoB predicted electronic properties polarizability (α), atomization energy (E), HOMO and LUMO eigenvalues (ϵ) for 2165 randomly drawn out-of-sample molecules from GDB-7 for training set sizes of $N = 1000$ and 5000, respectively.

The applicability and accuracy of the BoB model also extends for predicting properties other than energies, including polarizability and highest and lowest molecular orbital energies (HOMO, LUMO), all computed at the DFT-PBE0 level of theory. BoB error distributions for out-of-sample predictions are shown in Figure 6. For models trained on $N = 1000$ GDB-7 molecules with property data taken from Ref.,[22] the resulting MAE are 0.15 Å^3, 0.21 eV, 0.19 eV, for polarizability (mean of 11.11 Å^3), HOMO (mean of -7.02 eV), and LUMO (mean of -0.52 eV), respectively. For the $N = 5000$ BoB model, these respective errors reduce to 0.09 Å^3, 0.14 eV, and 0.12 eV. We remark that the BoB model once again performs as well as or better than the more complex ML models in the literature.[22]

The final question we would like to address is the feasibility of utilizing the BoB model in the context of high throughput calculations on molecular systems. This requires the assessment of the BoB model on a much larger dataset of molecules. To demonstrate that the power and robustness of our method extends beyond GDB-7, we employed the 134k dataset of quantum-chemical calculations (containing 133,885 molecules), recently presented in Ref.[23] Similar to the case of GDB-7 dataset, we obtain an accuracy of 2.0 kcal/mol for atomization energies in the 134k dataset when training the BoB model on 30% of the molecules.

In summary, we have devised and applied a systematic hierarchy of efficient models to estimate atomization energies and different electronic properties for a representative set of organic molecules. The developed BoB model is quite successful for non-equilibrium geometries, hinting that it could also be extended to study vibrational properties of molecules. In addition, the BoB model is demonstrated to be sufficiently robust as a tool in the context of high-throughput calculations throughout a representative subset of the chemical compound space.

Acknowledgement

We thank Dr. Matthias Rupp for inspiring discussions. This work is supported by the European Research Council (ERC-StG VDW-CMAT), DFG Grant No. MU 987/20, Natural Sciences and Engineering Research Council of Canada, by the BK21 program of NRF, the Einstein Foundation, and the Swiss National Science Foundation (Grant No. PP00P2_138932). This work used resources of the Argonne Leadership Computing Facility at Argonne National Laboratory, which is supported by the Office of Science of the U.S. DOE under Contract No. DE-AC02-06CH11357.

Supporting Information Available

Detailed description of the Bag-of-Bonds (BoB) model, kernel parameters, and pairwise potentials. This material is available free of charge via the Internet http://pubs.acs.org. This material is available free of charge via the Internet at `http://pubs.acs.org/`.

References

(1) Kirkpatrick, P.; Ellis, C. Chemical Space. *Nature* **2004**, *432*, 823.

(2) Schneider, G. Virtual Screening: An Endless Staircase? *Nature Rev.* **2010**, *9*, 273.

(3) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, 2009.

(4) Manzhos, S.; Carrington, T. Using Neural Networks to Represent Potential Surfaces as Sums of Products. *J. Chem. Phys.* **2006**, *125*, 194105.

(5) Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.

(6) Behler, J. Atom-Centered Symmetry Functions for Constructing High-Dimensional Neural Networks Potentials. *J. Chem. Phys.* **2011**, *134*, 074106.

(7) Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons. *Phys. Rev. Lett.* **2010**, *104*, 136403.

(8) Fletcher, T. L.; Davie, S. J.; Popelier, P. L. Prediction of Intramolecular Polarization of Aromatic Amino Acids using Kriging Machine Learning. *J. Chem. Theory and Comput.* **2014**, *10*, 3708–3719.

(9) Fink, T.; Bruggesser, H.; Reymond, J.-L. Virtual Exploration of the Small-Molecule Chemical Universe Below 160 Daltons. *Angew. Chem. Int. Ed.* **2005**, *44*, 1504.

(10) Fink, T.; Reymond, J.-L. Virtual Exploration of the Chemical Universe up to 11 Atoms of C, N, O, F: Assembly of 26.4 Million Structures (110.9 Million Stereoisomers) and Analysis for New Ring Systems, Stereochemistry, Physicochemical Properties, Compound Classes, and Drug Discovery. *J. Chem. Inf. Model.* **2007**, *47*, 342.

(11) Guha, R.; Howard, M. T.; Hutchison, G. R.; Murray-Rust, P.; Rzepa, H.; Steinbeck, C.; Wegner, J.; Willighagen, E. L. The Blue Obelisk – Interoperability in Chemical Informatics. *J. Chem. Inf. Model.* **2006**, *46*, 991.

(12) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865.

(13) Blum, V.; Gehrke, R.; Hanke, F.; Havu, P.; Havu, V.; Ren, X.; Reuter, K.; Scheffler, M. Ab Initio Molecular Simulations with Numeric Atom-Centered Orbitals. *Chem. Phys. Commun.* **2009**, *180*, 2175.

(14) Perdew, J. P.; Ernzerhof, M.; Burke, K. Rationale for Mixing Exact Exchange with Density Functional Approximations. *J. Chem. Phys.* **1996**, *105*, 9982.

(15) Lynch, B. J.; Truhlar, D. G. Robust and Affordable Multicoefficient Methods for Thermochemistry and Thermochemical. *J. Phys. Chem. A* **2003**, *107*, 3898.

(16) Hansen, K.; Montavon, G.; Biegler, F.; Fazli, S.; Rupp, M.; Scheffler, M.; von Lilienfeld, O. A.; Tkatchenko, A.; Müller, K.-R. Assessment and Validation of Machine Learning Methods for Predicting Molecular Atomization Energies. *J. Chem. Theory and Comput.* **2013**, *9*, 3404.

(17) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and Accurate

Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.

(18) Benson, S. W. III - Bond Energies. *J. Chem. Educ.* **1965**, *42*, 502.

(19) Forman, G. An Extensive Empirical Study of Feature Selection Metrics for Text Classification. *The Journal of Machine Learning Research* **2003**, *3*, 1289.

(20) Joachims, T. *Text categorization with support vector machines: Learning with many relevant features*; Springer, 1998.

(21) Patterson, A. L. Homometric Structures. *Nature* **1939**, *143*, 939.

(22) Montavon, G.; Rupp, M.; Gobre, V.; Vazquez-Mayagoitia, A.; Hansen, K.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Machine Learning of Molecular Electronic Properties in Chemical Compound Space. *New J. Phys.* **2013**, *15*, 095003.

(23) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum Chemistry Structures and Properties of 134 Kilo Molecules. *Sci. Data* **2014**, *1*, 140022.